Chapter 2

Ordinary Least Squares

The bread and butter of regression analysis is the estimation of the coefficients of econometric models using a technique called Ordinary Least Squares (OLS). The first two sections of this chapter summarize the reasoning behind and the mechanics of OLS. Regression users rely on computers to do the actual OLS calculations, so the emphasis here is on understanding what OLS attempts to do and how it goes about doing it.

How can you tell a good equation from a bad one once it has been estimated? There are a number of useful criteria, including the extent to which the estimated equation fits the actual data. A focus on fit is not without perils, however, so we share an example of the misuse of this criterion.

The chapter concludes with a new kind of learning tool that we call an econometric lab.

2.1 Estimating Single-Independent-Variable Models with OLS

The purpose of regression analysis is to take a purely theoretical equation like:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (2.1)$$

and use a set of data to create an estimated equation like:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{2.2}$$

where each "hat" indicates a sample estimate of the true population value. (In the case of Y, the "true population value" is $E[Y|X]$.) The purpose of the estimation technique is to obtain numerical values for the coefficients of an otherwise completely theoretical regression equation.

The most widely used method of obtaining these estimates is Ordinary Least Squares (OLS), which has become so standard that its estimates are presented as a point of reference even when results from other estimation techniques are used. **Ordinary Least Squares (OLS)** is a regression estimation technique that calculates the $\hat{\beta}$s so as to minimize the sum of the squared residuals, thus:[1]

$$\text{OLS minimizes } \sum_{i=1}^{N} e_i^2 \qquad (i = 1, 2, \ldots, N) \tag{2.3}$$

Since these residuals (e_is) are the differences between the actual Ys and the estimated Ys produced by the regression (the $\hat{Y}$s in Equation 2.2), Equation 2.3 is equivalent to saying that OLS minimizes $\sum (Y_i - \hat{Y}_i)^2$.

Why Use Ordinary Least Squares?

Although OLS is the most-used regression estimation technique, it's not the only one. Indeed, econometricians have developed what seem like zillions of different estimation techniques, a number of which we'll discuss later in this text.

There are at least three important reasons for using OLS to estimate regression models:

1. OLS is relatively easy to use.
2. The goal of minimizing $\sum e_i^2$ is quite appropriate from a theoretical point of view.
3. OLS estimates have a number of useful characteristics.

[1] The summation symbol, $\sum$, indicates that all terms to its right should be added (or summed) over the range of the *i* values attached to the bottom and top of the symbol. In Equation 2.3, for example, this would mean adding up e_i^2 for all integer values between 1 and N:

$$\sum_{i=1}^{N} e_i^2 = e_1^2 + e_2^2 + \cdots + e_N^2$$

Often the $\sum$ notation is simply written as $\sum_i$, and it is assumed that the summation is over all observations from $i = 1$ to $i = N$. Sometimes, the i is omitted entirely and the same assumption is made implicitly.

The first reason for using OLS is that it's the simplest of all econometric estimation techniques. Most other techniques involve complicated nonlinear formulas or iterative procedures, many of which are extensions of OLS itself. In contrast, OLS estimates are simple enough that, if you had to, you could calculate them without using a computer or a calculator (for a single-independent-variable model). Indeed, in the "dark ages" before computers and calculators, econometricians calculated OLS estimates by hand!

The second reason for using OLS is that minimizing the summed, squared residuals is a reasonable goal for an estimation technique. To see this, recall that the residual measures how close the estimated regression equation comes to the actual observed data:

$$e_i = Y_i - \hat{Y}_i \qquad (i = 1, 2, \ldots, N) \tag{1.15}$$

Since it's reasonable to want our estimated regression equation to be as close as possible to the observed data, you might think that you'd want to minimize these residuals. The main problem with simply totaling the residuals is that e_i can be negative as well as positive. Thus, negative and positive residuals might cancel each other out, allowing a wildly inaccurate equation to have a very low $\sum e_i$. For example, if Y = 100,000 for two consecutive observations and if your equation predicts 1.1 million and −900,000, respectively, your residuals will be +1 million and −1 million, which add up to zero!

We could get around this problem by minimizing the sum of the absolute values of the residuals, but absolute values are difficult to work with mathematically. Luckily, minimizing the summed squared residuals does the job. Squared functions pose no unusual mathematical difficulties in terms of manipulations, and the technique avoids canceling positive and negative residuals because squared terms are always positive.

The final reason for using OLS is that its estimates have at least two useful properties:[2]

1. The sum of the residuals is exactly zero.
2. OLS can be shown to be the "best" estimator possible under a set of specific assumptions. We'll define "best" in Chapter 4.

An **estimator** is a mathematical technique that is applied to a sample of data to produce a real-world numerical **estimate** of the true population regression coefficient (or other parameters). Thus, OLS is an estimator, and a $\hat{\beta}$ produced by OLS is an estimate.

[2] These properties, and indeed all the properties of OLS that we discuss in this book, are true as long as a constant term is included in the regression equation. For more on this, see Section 7.1.

How Does OLS Work?

How would OLS estimate a single-independent-variable regression model like Equation 2.1?

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2.1}$$

OLS selects those estimates of β_0 and β_1 that minimize the squared residuals, summed over all the sample data points.

For an equation with just one independent variable, these coefficients are:[3]

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} [(X_i - \overline{X})(Y_i - \overline{Y})]}{\sum_{i=1}^{N} (X_i - \overline{X})^2} \tag{2.4}$$

and, given this estimate of β_1,

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{2.5}$$

where $\overline{X}$ = the mean of X, or $\sum X_i/N$, and $\overline{Y}$ = the mean of Y, or $\sum Y_i/N$. Note that for each different data set, we'll get different estimates of β_1 and β_0, depending on the sample.

[3] Since

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

and $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta} X_{1i}$, OLS actually minimizes

$$\sum_i e_i^2 = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

by choosing the $\hat{\beta}$s that do so. For those with a moderate grasp of calculus and algebra, the derivation of these equations is informative.

An Illustration of OLS Estimation

The equations for calculating regression coefficients might seem a little forbidding, but it's not hard to apply them yourself to data sets that have only a few observations and independent variables. Although you'll usually want to use regression software packages to do your estimation, you'll understand OLS better if you work through the following illustration.

To keep things simple, let's attempt to estimate the regression coefficients of the height and weight data given in Section 1.4. For your convenience in following this illustration, the original data are reproduced in Table 2.1. As was noted previously, the formulas for OLS estimation for a regression equation with one independent variable are Equations 2.4 and 2.5:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}[(X_i - \overline{X})(Y_i - \overline{Y})]}{\sum_{i=1}^{N}(X_i - \overline{X})^2} \tag{2.4}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} \tag{2.5}$$

If we undertake the calculations outlined in Table 2.1 and substitute them into Equations 2.4 and 2.5, we obtain these values:

$$\hat{\beta}_1 = \frac{590.20}{92.50} = 6.38$$

$$\hat{\beta}_0 = 169.4 - (6.38 \cdot 10.35) = 103.4$$

or

$$\hat{Y}_i = 103.4 + 6.38X_i \tag{2.6}$$

If you compare these estimates, you'll find that the manually calculated coefficient estimates are the same as the computer regression results summarized in Section 1.4.

As can be seen in Table 2.1, the sum of the $\hat{Y}$s (column 8) equals the sum of the Ys (column 2), so the sum of the residuals (column 9) does indeed equal zero (except for rounding errors).

Table 2.1 The Calculation of Estimated Regression Coefficients for the Weight/Height Example

Raw Data			Required Intermediate Calculations					
i (1)	Y_i (2)	X_i (3)	$(Y_i - \bar{Y})$ (4)	$(X_i - \bar{X})$ (5)	$(X_i - \bar{X})^2$ (6)	$(X_i - \bar{X})(Y_i - \bar{Y})$ (7)	$\hat{Y}_i$ (8)	$e_i = Y_i - \hat{Y}_i$ (9)
1	140	5	−29.40	−5.35	28.62	157.29	135.3	4.7
2	157	9	−12.40	−1.35	1.82	16.74	160.8	−3.8
3	205	13	35.60	2.65	7.02	94.34	186.3	18.7
4	198	12	28.60	1.65	2.72	47.19	179.9	18.1
5	162	10	−7.40	−0.35	0.12	2.59	167.2	−5.2
6	174	11	4.60	0.65	0.42	2.99	173.5	0.5
7	150	8	−19.40	−2.35	5.52	45.59	154.4	−4.4
8	165	9	−4.40	−1.35	1.82	5.94	160.8	4.2
9	170	10	0.60	−0.35	0.12	−0.21	167.2	2.8
10	180	12	10.60	1.65	2.72	17.49	179.9	0.1
11	170	11	0.60	0.65	0.42	0.39	173.5	−3.5
12	162	9	−7.40	−1.35	1.82	9.99	160.8	1.2
13	165	10	−4.40	−0.35	0.12	1.54	167.2	−2.2
14	180	12	10.60	1.65	2.72	17.49	179.9	0.1
15	160	8	−9.40	−2.35	5.52	22.09	154.4	5.6
16	155	9	−14.40	−1.35	1.82	19.44	160.8	−5.8
17	165	10	−4.40	−0.35	0.12	1.54	167.2	−2.2
18	190	15	20.60	4.65	21.62	95.79	199.1	−9.1
19	185	13	15.60	2.65	7.02	41.34	186.3	−1.3
20	155	11	−14.40	0.65	0.42	−9.36	173.5	−18.5
Sum	3388	207	0.0	0.0	92.50	590.20	3388.0	−0.0
Mean	169.4	10.35	0.0	0.0			169.4	0.0

2.2 Estimating Multivariate Regression Models with OLS

Let's face it: only a few dependent variables can be explained fully by a single independent variable. A person's weight, for example, is influenced by more than just that person's height. What about bone structure, percent body fat, exercise habits, or diet?

As important as additional explanatory variables might seem to the height/weight example, there's even more reason to include a variety of independent variables in economic and business applications. Although the per capita quantity consumed of a product is certainly affected by price, that's not the whole story. Advertising, per capita income, the prices of substitutes, the influence of foreign markets, the quality of customer service, possible fads, and changing tastes all are important in real-world models. As a result, it's vital to move from single-independent-variable regressions to *multivariate regression models,* or equations with more than one independent variable.

The Meaning of Multivariate Regression Coefficients

The general multivariate regression model with K independent variables can be represented by Equation 1.11:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i \qquad (1.11)$$

where i, as before, goes from 1 to N and indicates the observation number. Thus, X_{1i} indicates the *i*th observation of independent variable X_1, while X_{2i} indicates the *i*th observation of another independent variable, X_2.

The biggest difference between a single-independent-variable regression model and a multivariate regression model is in the interpretation of the latter's slope coefficients. These coefficients, often called *partial* regression coefficients, are defined to allow a researcher to distinguish the impact of one variable from that of other independent variables.

Specifically, a **multivariate regression coefficient** indicates the change in the dependent variable associated with a one-unit increase in the independent variable in question, *holding constant the other independent variables in the equation.*

This last italicized phrase is a key to understanding multiple regression (as multivariate regression is often called). The coefficient β_1 measures the impact on Y of a one-unit increase in X_1, holding constant $X_2, X_3, \ldots$ and X_K but *not* holding constant any relevant variables that might have been omitted from the equation (e.g., X_{K+1}). The coefficient β_0 is the value of Y when all the Xs and the error term equal zero. As you'll learn in Section 7.1, you should always include a constant term in a regression equation, but you should not rely on estimates of β_0 for inference.

As an example, let's consider the following annual model of the per capita consumption of beef in the United States:

$$\widehat{CB}_t = 37.54 - 0.88P_t + 11.9Yd_t \qquad (2.7)$$

where: CB_t = the per capita consumption of beef in year t (in pounds per person)
P_t = the price of beef in year t (in cents per pound)
Yd_t = the per capita disposable income in year t (in thousands of dollars)

The estimated coefficient of income, 11.9, tells us that beef consumption will increase by 11.9 pounds per person if per capita disposable income goes up by $1,000, holding constant the price of beef. The ability to hold price constant is crucial because we'd expect such a large increase in per capita income to stimulate demand, therefore pushing up prices and making it hard to distinguish the effect of the income increase from the effect of the price increase. The multivariate regression estimate allows us to focus on the impact of the income variable by holding the price variable constant. Note, however, that the equation does not hold constant other possible variables (like the price of a substitute) because these variables are not included in Equation 2.7.

Before we move on to the next section, let's take the time to analyze the estimated coefficients of Equation 2.7 in a bit more depth. First, the coefficient of P tells us the impact of a one-cent increase in the price of beef on the per capita consumption of beef, holding constant per capita income. Do you agree that the estimated coefficient has the sign that economic theory would predict? Second, think about how the estimated coefficients would change if we were to change the units of disposable income from "thousands of dollars" to "dollars." The estimated equation would remain the same except that the coefficient of Yd would decrease from 11.9 to 0.0119.

OLS Estimation of Multivariate Regression Models

The application of OLS to an equation with more than one independent variable is quite similar to its application to a single-independent-variable model. To see this, consider the estimation of the simplest possible multivariate model, one with just two independent variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{2.8}$$

The goal of OLS is to choose those $\hat{\beta}$s that minimize the summed squared residuals. These residuals are now from a multivariate model, but they can be minimized using the same mathematical approach used in Section 2.1. Thus the OLS estimation of multivariate models is identical in general approach to the OLS estimation of models with just one independent variable. The equations themselves are more cumbersome,[4] but the underlying principle of estimating $\hat{\beta}$s that minimize the summed squared residuals remains the same.

Luckily, user-friendly computer packages can calculate estimates with these unwieldy equations in less than a second of computer time. Indeed, only someone lost in time or stranded on a desert island would bother estimating a multivariate regression model without a computer. The rest of us will use Stata, EViews, SPSS, SAS, or any of the other commercially available regression packages.

An Example of a Multivariate Regression Model

As an example of multivariate regression, let's take a look at a model of financial aid awards at a liberal arts college. The dependent variable in such a study would be the amount, in dollars, awarded to a particular financial aid applicant:

FINAID_i = the financial aid (measured in dollars of grant per year) awarded to the *i*th applicant

What kinds of independent variables might influence the amount of financial aid received by a given student? Well, most aid is either need-based or merit-based, so it makes sense to consider a model that includes at least these two attributes:

$$\text{FINAID}_i = \beta_0 + \overset{-}{\beta_1}\text{PARENT}_i + \overset{+}{\beta_2}\text{HSRANK}_i + \epsilon_i \tag{2.9}$$

where: PARENT_i = the amount (in dollars per year) that the parents of the *i*th student are judged able to contribute to college expenses

HSRANK_i = the *i*th student's GPA rank in high school, measured as a percentage (ranging from a low of 0 to a high of 100)

[4] For Equation 2.8, the estimated coefficients are:

$$\hat{\beta}_1 = \frac{(\sum yx_1)(\sum x_2^2) - (\sum yx_2)(\sum x_1x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

$$\hat{\beta}_2 = \frac{(\sum yx_2)(\sum x_1^2) - (\sum yx_1)(\sum x_1x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}_1 - \hat{\beta}_2\overline{X}_2$$

where lowercase variables indicate deviations from the mean, as in $y = Y_i - \overline{Y}$; $x_1 = X_{1i} - \overline{X}_1$; and $x_2 = X_{2i} - \overline{X}_2$.

Note from the signs over the coefficients in Equation 2.9 that we anticipate that the more parents can contribute to their child's education, the less the financial aid award will be. Similarly, we expect that the higher the student's rank in high school, the higher the financial aid award will be. Do you agree with these expectations?

If we estimate Equation 2.9 using OLS and the data[5] in Table 2.2, we get:

$$\widehat{\text{FINAID}}_i = 8927 - 0.36\text{PARENT}_i + 87.4\text{HSRANK}_i \qquad (2.10)$$

What do these coefficients mean? Well, the –0.36 means that the model implies that the *i*th student's financial aid grant will fall by $0.36 for every dollar increase in his or her parents' ability to pay, holding constant high school rank. Does the sign of the estimated coefficient meet our expectations? Yes. Does the size of the coefficient make sense? Yes.

To be sure that you understand this concept, take the time to write down the meaning of the coefficient of HSRANK in Equation 2.10. Do you agree that the model implies that the *i*th student's financial aid grant will increase by $87.40 for each percentage point increase in high school rank, holding constant parents' ability to pay? Does this estimated coefficient seem reasonable?

Let's analyze Equation 2.10. Suppose someone told you that they believed that HSRANK is the most important variable in the model because its coefficient, 87.4, is much larger than the coefficient of FINAID. Would you agree? Before you answer, consider what Equation 2.10 would look like if the units of measurement of PARENT had been thousands of dollars instead of dollars:

$$\widehat{\text{FINAID}} = 8927 - 357\text{PARENT} + 87.4\text{HSRANK} \qquad (2.11)$$

Whoops! That puts things in a different light. Now the coefficient of PARENT is much larger than the coefficient of HSRANK. Since the size of a coefficient clearly depends on the units of measurement of the variable, we can't use coefficient size alone to make judgments about the importance of a variable. For more on this issue, see Section 5.4.

Take a look at Figures 2.1 and 2.2. These figures contain two different views of Equation 2.10. Figure 2.1 is a diagram of the effect of PARENT on FINAID, holding HSRANK constant, and Figure 2.2 shows the effect of HSRANK on FINAID, holding PARENT constant. These two figures are graphical representations of multivariate regression coefficients, since they measure the impact

[5] These data are from an unpublished analysis of financial aid awards at Occidental College. The fourth variable in Table 2.2 is MALE_i, which equals 1 if the *i*th student is male and 0 otherwise.

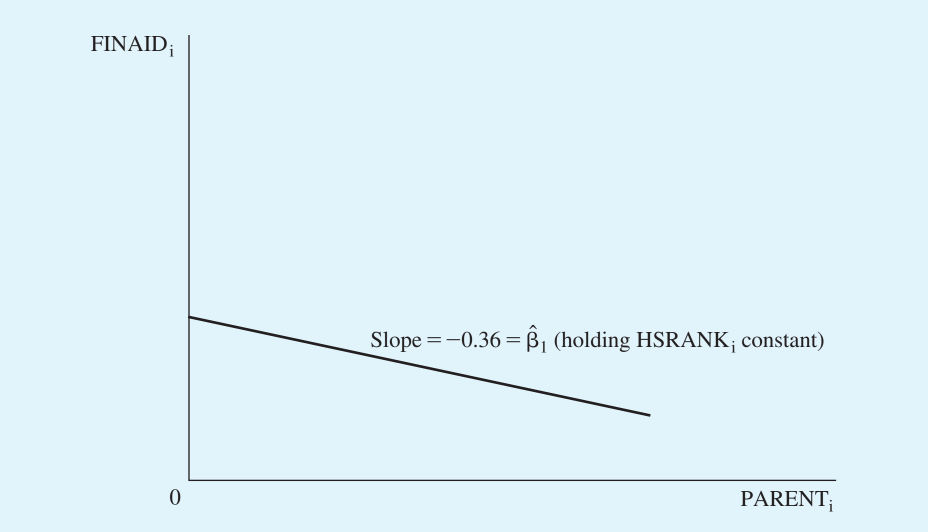

Figure 2.1 Financial Aid as a Function of Parents' Ability to Pay

In Equation 2.10, an increase of one dollar in the parents' ability to pay decreases the financial aid award by $0.36, holding constant high school rank.

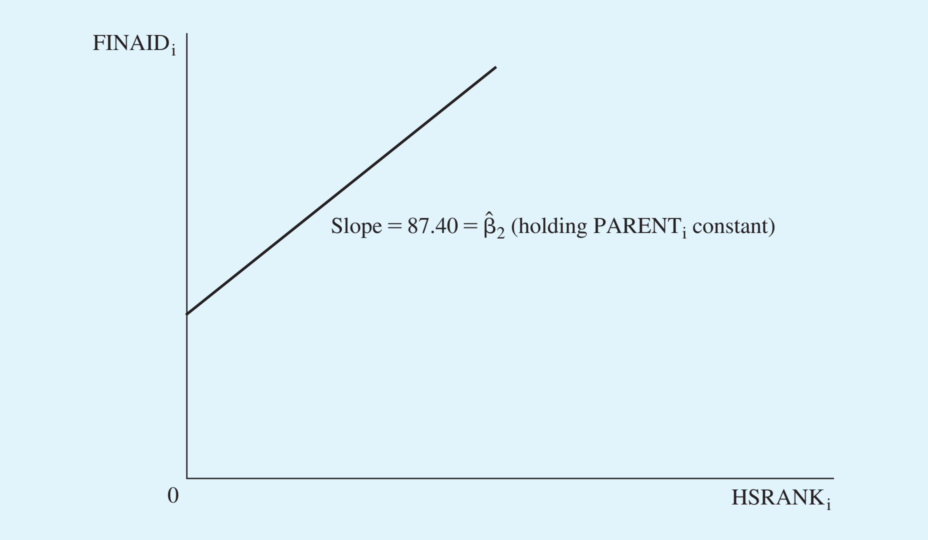

Figure 2.2 Financial Aid as a Function of High School Rank

In Equation 2.10, an increase of one percentage point in high school rank increases the financial aid award by $87.40, holding constant parents' ability to pay.

Table 2.2 Data for the Financial Aid Example

i	FINAID	PARENT	HSRANK	MALE
1	19,640	0	92	0
2	8,325	9,147	44	1
3	12,950	7,063	89	0
4	700	33,344	97	1
5	7,000	20,497	95	1
6	11,325	10,487	96	0
7	19,165	519	98	1
8	7,000	31,758	70	0
9	7,925	16,358	49	0
10	11,475	10,495	80	0
11	18,790	0	90	0
12	8,890	18,304	75	1
13	17,590	2,059	91	1
14	17,765	0	81	0
15	14,100	15,602	98	0
16	18,965	0	80	0
17	4,500	22,259	90	1
18	7,950	5,014	82	1
19	7,000	34,266	98	1
20	7,275	11,569	50	0
21	8,000	30,260	98	1
22	4,290	19,617	40	1
23	8,175	12,934	49	1
24	11,350	8,349	91	0
25	15,325	5,392	82	1
26	22,148	0	98	0
27	17,420	3,207	99	0
28	18,990	0	90	0
29	11,175	10,894	97	0
30	14,100	5,010	59	0
31	7,000	24,718	97	1
32	7,850	9,715	84	1
33	0	64,305	84	0
34	7,000	31,947	98	1
35	16,100	8,683	95	1
36	8,000	24,817	99	0
37	8,500	8,720	20	1
38	7,575	12,750	89	1
39	13,750	2,417	41	1
40	7,000	26,846	92	1
41	11,200	7,013	86	1
42	14,450	6,300	87	0

Table 2.2 (*continued*)

i	FINAID	PARENT	HSRANK	MALE
43	15,265	3,909	84	0
44	20,470	2,027	99	1
45	9,550	12,592	89	0
46	15,970	0	57	0
47	12,190	6,249	84	0
48	11,800	6,237	81	0
49	21,640	0	99	0
50	9,200	10,535	68	0

Datafile = FINAID2

on the dependent variable of a given independent variable, holding constant the other variables in the equation.

Total, Explained, and Residual Sums of Squares

Before going on, let's pause to develop some measures of how much of the variation of the dependent variable is explained by the estimated regression equation. Such comparison of the estimated values with the actual values can help a researcher judge the adequacy of an estimated regression.

Econometricians use the squared variations of Y around its mean as a measure of the amount of variation to be explained by the regression. This computed quantity is usually called the **total sum of squares**, or TSS, and is written as:

$$\mathrm{TSS} = \sum_{i=1}^{N} (Y_i - \overline{Y})^2 \tag{2.12}$$

For Ordinary Least Squares, the total sum of squares has two components, variation that can be explained by the regression and variation that cannot:

$$\sum_i (Y_i - \overline{Y})^2 = \sum_i (\hat{Y}_i - \overline{Y})^2 + \sum_i e_i^2 \tag{2.13}$$

$$\begin{matrix} \text{Total Sum} & = & \text{Explained} & + & \text{Residual} \\ \text{of} & & \text{Sum of} & & \text{Sum of} \\ \text{Squares} & & \text{Squares} & & \text{Squares} \\ \text{(TSS)} & & \text{(ESS)} & & \text{(RSS)} \end{matrix}$$

This is usually called the *decomposition of variance*.

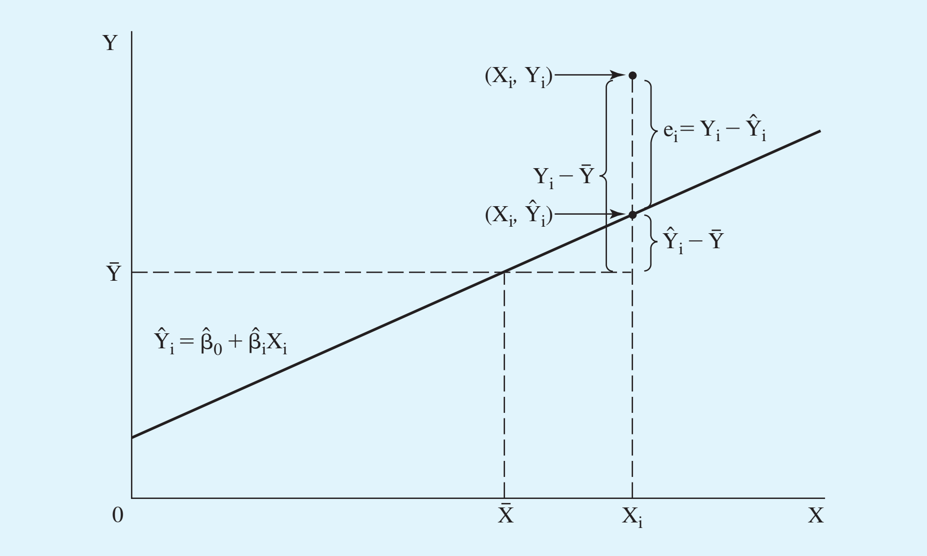

Figure 2.3 Decomposition of the Variance in Y

The variation of Y around its mean $(Y - \bar{Y})$ can be decomposed into two parts: (1) $(\hat{Y}_i - \bar{Y})$, the difference between the estimated value of $Y(\hat{Y})$ and the mean value of Y $(\bar{Y})$; and (2) $(Y_i - \hat{Y}_i)$, the difference between the actual value of Y and the estimated value of Y.

Figure 2.3 illustrates the decomposition of variance for a simple regression model. The estimated values of Y_i lie on the estimated regression line $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. The variation of Y around its mean $(Y_i - \bar{Y})$ can be decomposed into two parts: (1) $(\hat{Y}_i - \bar{Y})$, the difference between the estimated value of Y $(\hat{Y})$ and the mean value of Y $(\bar{Y})$; and (2) $(Y_i - \hat{Y}_i)$, the difference between the actual value of Y and the estimated value of Y.

The first component of Equation 2.13 measures the amount of the squared deviation of Y_i from its mean that is explained by the regression line. This component of the total sum of the squared deviations, called the **explained sum of squares**, or ESS, is attributable to the fitted regression line. The unexplained portion of TSS (that is, unexplained in an empirical sense by the estimated regression equation), is called the **residual sum of squares**, or RSS.[6]

[6] Note that some authors reverse the definitions of RSS and ESS (defining ESS as $\sum e_i^2$), and other authors reverse the order of the letters, as in SSR.

We can see from Equation 2.13 that the smaller the RSS is relative to the TSS, the better the estimated regression line fits the data. OLS is the estimating technique that minimizes the RSS and therefore maximizes the ESS for a given TSS.

2.3 Evaluating the Quality of a Regression Equation

If the bread and butter of regression analysis is OLS estimation, then the heart and soul of econometrics is figuring out how good these OLS estimates are.

Many beginning econometricians have a tendency to accept regression estimates as they come out of a computer, or as they are published in an article, without thinking about the meaning or validity of those estimates. Such blind faith makes as much sense as buying an entire wardrobe of clothes without trying them on. Some of the clothes will fit just fine, but many others will turn out to be big (or small) mistakes.

Instead, the job of an econometrician is to carefully think about and evaluate every aspect of the equation, from the underlying theory to the quality of the data, before accepting a regression result as valid. In fact, most good econometricians spend quite a bit of time thinking about what to expect from an equation *before* they estimate that equation.

Once the computer estimates have been produced, however, it's time to evaluate the regression results. The list of questions that should be asked during such an evaluation is long. For example:

1. Is the equation supported by sound theory?
2. How well does the estimated regression fit the data?
3. Is the data set reasonably large and accurate?
4. Is OLS the best estimator to be used for this equation?
5. How well do the estimated coefficients correspond to the expectations developed by the researcher before the data were collected?
6. Are all the obviously important variables included in the equation?
7. Has the most theoretically logical functional form been used?
8. Does the regression appear to be free of major econometric problems?

The goal of this text is to help you develop the ability to ask and appropriately answer these kinds of questions. In fact, the number in front of each question above roughly corresponds to the chapter in which we'll address the issues raised by that question. Since this is Chapter 2, it'll come as no surprise to you to hear that the rest of the chapter will be devoted to the second of these topics—the overall fit of the estimated model.

2.4 Describing the Overall Fit of the Estimated Model

Let's face it: we expect that a good estimated regression equation will explain the variation of the dependent variable in the sample fairly accurately. If it does, we say that the estimated model fits the data well.

Looking at the overall fit of an estimated model is useful not only for evaluating the quality of the regression, but also for comparing models that have different data sets or combinations of independent variables. We can never be sure that one estimated model represents the truth any more than another, but evaluating the quality of the fit of the equation is one ingredient in a choice between different formulations of a regression model. Be careful, however! The quality of the fit is a minor ingredient in this choice, and many beginning researchers allow themselves to be overly influenced by it.

R^2

The simplest commonly used measure of fit is R^2, or the coefficient of determination. **R^2** is the ratio of the explained sum of squares to the total sum of squares:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2} \qquad (2.14)$$

The higher R^2 is, the closer the estimated regression equation fits the sample data. Measures of this type are called "goodness of fit" measures. R^2 measures the percentage of the variation of Y around $\bar{Y}$ that is explained by the regression equation. Since OLS selects the coefficient estimates that minimize RSS, OLS provides the largest possible R^2, given a linear model. Since TSS, RSS, and ESS are all nonnegative (being squared deviations), and since ESS $\leq$ TSS, then R^2 must lie in the interval $0 \leq R^2 \leq 1$. A value of R^2 close to one shows an excellent overall fit, whereas a value near zero shows a failure of the estimated regression equation to explain the values of Y_i better than could be explained by the sample mean $\bar{Y}$.

Figures 2.4 through 2.6 demonstrate some extremes. Figure 2.4 shows an X and Y that are unrelated. The fitted regression line might as well be $\hat{Y} = \bar{Y}$, the same value it would have if X were omitted. As a result, the estimated

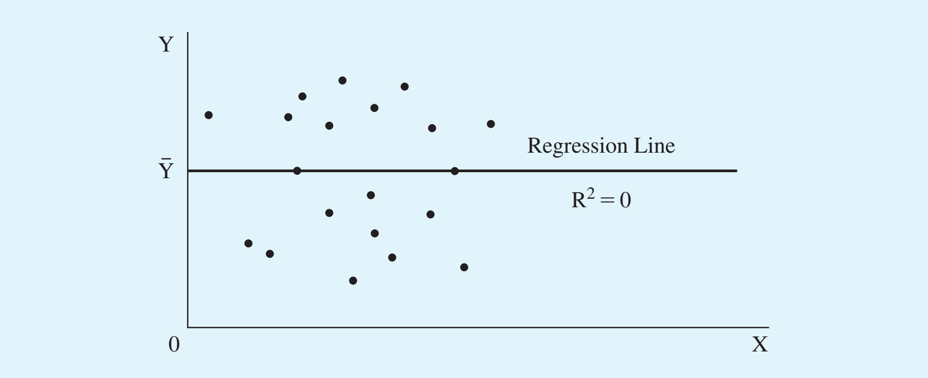

Figure 2.4

X and Y are not related; in such a case, R^2 would be 0.

linear regression is no better than the sample mean as an estimate of Y_i. The explained portion, ESS, = 0, and the unexplained portion, RSS, equals the total squared deviations TSS; thus, $R^2 = 0$.

Figure 2.5 shows a relationship between X and Y that can be "explained" quite well by a linear regression equation: the value of R^2 is .95. This kind of result is typical of a time-series regression with a good fit. Most of the variation has been explained, but there still remains a portion of the variation that is essentially random or unexplained by the model.

Goodness of fit is relative to the topic being studied. In time series data, we often get a very high R^2 because there can be significant time trends on both

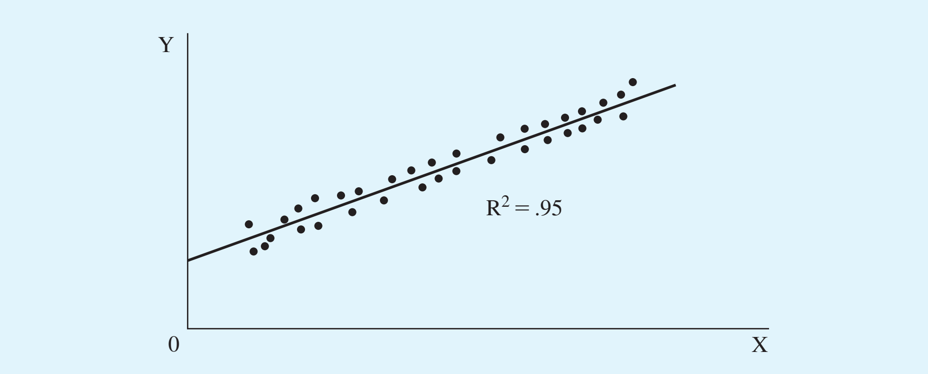

Figure 2.5

A set of data for X and Y that can be "explained" quite well with a regression line ($R^2 = .95$).

sides of the equation. In cross-sectional data, we often get low R^2s because the observations (say, countries) differ in ways that are not easily quantified. In such a situation, an R^2 of .50 might be considered a good fit, and researchers would tend to focus on identifying the variables that have a substantive impact on the dependent variable, not on R^2. In other words, there is no simple method of determining how high R^2 must be for the fit to be considered satisfactory. Instead, knowing when R^2 is relatively large or small is a matter of experience. It should be noted that a high R^2 does not imply that changes in X lead to changes in Y, as there may be an underlying variable whose changes lead to changes in both X and Y simultaneously.

Figure 2.6 shows a perfect fit of $R^2 = 1$. Such a fit implies that no estimation is required. The relationship is completely deterministic, and the slope and intercept can be calculated from the coordinates of any two points. In fact, reported equations with R^2s equal to (or very near) one should be viewed with suspicion; they very likely do not explain the movements of the dependent variable Y in terms of the causal proposition advanced, even though they explain them empirically. This caution applies to economic applications, but not necessarily to those in fields like physics or chemistry.

$\bar{R}^2$, The Adjusted R^2

A major problem with R^2 is that adding another independent variable to a particular equation can never decrease R^2. That is, if you compare two equations that are identical (same dependent variable and independent variables), except that one has an additional independent variable, the equation

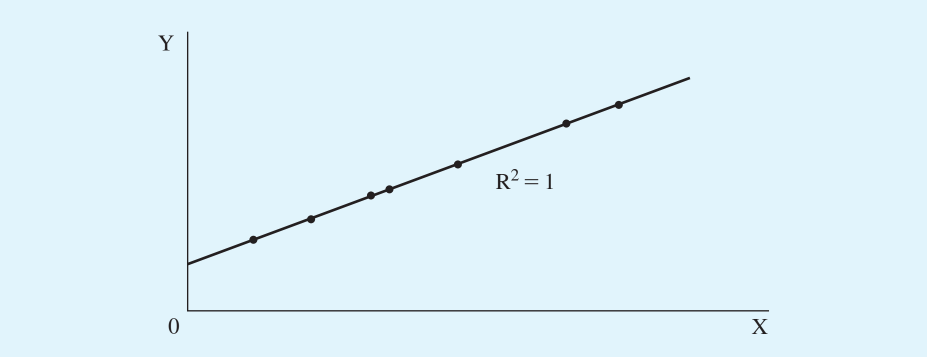

Figure 2.6

A perfect fit: all the data points are on the regression line, and the resulting R^2 is 1.

with the greater number of independent variables will always have a better (or equal) fit as measured by R^2.

To see this, recall the equation for R^2, Equation 2.14.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2} \tag{2.14}$$

What will happen to R^2 if we add a variable to the equation? Adding a variable can't change TSS (can you figure out why?), but in most cases the added variable will reduce RSS, so R^2 will rise. You know that RSS will never increase because the OLS program could always set the coefficient of the added variable equal to zero, thus giving the same fit as the previous equation. The coefficient of the newly added variable being zero is the only circumstance in which R^2 will stay the same when a variable is added. Otherwise, R^2 will always increase when a variable is added to an equation.

Perhaps an example will make this clear. Let's return to our weight guessing regression, Equation 1.19:

$$\text{Estimated weight} = 103.40 + 6.38\ \text{Height (over five feet)}$$

The R^2 for this equation is .74. If we now add a completely nonsensical variable to the equation (say, the campus post office box number of each individual in question), then it turns out that the results become:

$$\text{Estimated weight} = 102.35 + 6.36\ (\text{Height} > \text{five feet}) + 0.02\ (\text{Box\#})$$

but the R^2 for this equation is .75! Thus, an individual using R^2 alone as the measure of the quality of the fit of the regression would choose the second version as better fitting.

The inclusion of the campus post office box variable not only adds a nonsensical variable to the equation, but it also requires the estimation of another coefficient. This lessens the **degrees of freedom**, or the excess of the number of observations (N) over the number of coefficients (including the intercept) estimated (K + 1). For instance, when the campus box number variable is added to the weight/height example, the number of observations stays constant at 20, but the number of estimated coefficients increases from 2 to 3, so the number of degrees of freedom falls from 18 to 17. This decrease has a cost, since the lower the degrees of freedom, the less reliable the estimates are likely to be. Thus, the increase in the quality of the fit caused by the addition of a variable needs to be compared to the decrease in the degrees of freedom before a decision can be made with respect to the statistical impact of the added variable.

To sum, R^2 is of little help if we're trying to decide whether adding a variable to an equation improves our ability to meaningfully explain the dependent variable. Because of this problem, econometricians have developed another measure of the quality of the fit of an equation. That measure is $\bar{\mathbf{R}}^2$ (pronounced R-bar-squared), which is R^2 adjusted for degrees of freedom:

$$\bar{R}^2 = 1 - \frac{\sum e_i^2/(N - K - 1)}{\sum (Y_i - \bar{Y})^2/(N - 1)} \tag{2.15}$$

$\bar{R}^2$ measures the percentage of the variation of Y around its mean that is explained by the regression equation, *adjusted for degrees of freedom.*

$\bar{R}^2$ will increase, decrease, or stay the same when a variable is added to an equation, depending on whether the improvement in fit caused by the addition of the new variable outweighs the loss of the degree of freedom. An increase in $\bar{R}^2$ indicates that the marginal benefit of adding a variable exceeds the cost, while a decrease in $\bar{R}^2$ indicates that the marginal cost exceeds the benefit. Indeed, the $\bar{R}^2$ for the weight-guessing equation *decreases* to .72 when the mail box variable is added. The mail box variable, since it has no theoretical relation to weight, should never have been included in the equation, and the $\bar{R}^2$ measure supports this conclusion.

The highest possible $\bar{R}^2$ is 1.00, the same as for R^2. The lowest possible $\bar{R}^2$, however, is not .00; if R^2 is extremely low, $\bar{R}^2$ can be slightly negative.

$\bar{R}^2$ can be used to compare the fits of equations with the same dependent variable and different numbers of independent variables. Because of this property, most researchers automatically use $\bar{R}^2$ instead of R^2 when evaluating the fit of their estimated regression equations. Note, however, that $\bar{R}^2$ is not as useful when comparing the fits of two equations that have different dependent variables or dependent variables that are measured differently.

Finally, a warning is in order. Always remember that the quality of fit of an estimated equation is only one measure of the overall quality of that regression. As mentioned previously, the degree to which the estimated coefficients conform to economic theory and the researcher's previous expectations about those coefficients are just as important as the fit itself. For instance, an estimated equation with a good fit but with an implausible sign for an estimated coefficient might give implausible predictions and thus not be a very useful equation. Other factors, such as theoretical relevance and usefulness, also come into play. Let's look at an example of these factors.

2.5 An Example of the Misuse of $\bar{R}^2$

Section 2.4 implies that the higher the overall fit of a given equation, the better. Unfortunately, many beginning researchers assume that if a high $\bar{R}^2$ is good, then maximizing $\bar{R}^2$ is the best way to maximize the quality of an equation. Such an assumption is dangerous because a good overall fit is only one measure of the quality of an equation.

Perhaps the best way to visualize the dangers inherent in maximizing $\bar{R}^2$ without regard to the economic meaning or statistical significance of an equation is to look at an example of such misuse. This is important because it is one thing for a researcher to agree in theory that "$\bar{R}^2$ maximizing" is bad, and it is another thing entirely for that researcher to avoid subconsciously maximizing $\bar{R}^2$ on projects. It is easy to agree that the goal of regression is not to maximize $\bar{R}^2$, but many researchers find it hard to resist that temptation.

As an example, suppose you decide to combine your love of pizza with your love of economics, and think it might be a good idea to estimate a model of the determinants of mozzarella cheese consumption. You do some research and learn that mozzarella is a normal good, so you include income in your model. You collect a small sample, estimate the equation, and get the following results:

$$\widehat{\text{MOZZARELLA}_t} = -0.85 + 0.378\text{INCOME}_t \qquad (2.16)$$

$$N = 10 \qquad \bar{R}^2 = .88$$

where: MOZZARELLA_t = U.S. per capita consumption of mozzarella cheese (in pounds) in year t

INCOME_t = U.S. real disposable per capita income (in thousands of dollars) in year t

You think, "What a great fit!" But like many budding econometricians, you wonder . . . could you do even better by adding another independent variable? You find some interesting data and, on a hunch, add a variable to the model and re-run the regression:[7]

$$\widehat{\text{MOZZARELLA}_t} = 3.33 + 0.248\text{INCOME}_t - 0.046\text{DROWNINGS}_t \qquad (2.17)$$

$$N = 10 \qquad \bar{R}^2 = .97$$

where: DROWNINGS_t = U.S. deaths due to drowning after falling out of a fishing boat in year t

[7] This equation was created by Bruce Johnson of Centre College for the years 2000–2009. The data from MOZZARELLA and DROWNINGS come from tylervigen.com, while the data for INCOME are from the 2011 *Economic Report of the President*, Table B-31. For more examples of this type, see Tyler Vigen, *Spurious Correlation* (New York: Hachette Books, 2015).

The second equation has a much higher $\bar{R}^2$ than the first one, which you chose on the basis of theory. Does this mean your second model is better? Before you answer, reread the warning at the bottom of page 54 about the quality of fit $(\bar{R}^2)$ being only one measure of the overall quality of a regression.

Did you read it? OK, then you probably know that the answer is . . . NO! Equation 2.17 has a better fit, but it's preposterous to think that the number of drownings belongs in an equation for per capita mozzarella cheese consumption. No reasonable economic theory could link drownings to cheese consumption! What happened here is that in this small sample, DROWNINGS is highly correlated with MOZZARELLA. For no particular reason other than coincidence, drowning deaths from falling out of fishing boats went steadily down from 2000 to 2009, while per capita consumption of mozzarella cheese went up, so adding DROWNINGS boosted $\bar{R}^2$. This increased fit, however, doesn't mean that the equation is better. Such a meaningless result is called a *spurious regression*[8] and it should be ignored. In all honesty, Equation 2.17 never should have been run in the first place. It's too fishy an idea to take seriously.

Thus, a researcher who uses $\bar{R}^2$ as the sole measure of the quality of an equation (at the expense of economic theory or statistical significance) increases the chances of having unrepresentative or misleading results. This practice should be avoided at all costs. No simple rule of econometric estimation is likely to work in all cases. Instead, a combination of technical competence, theoretical judgment, and common sense makes for a good econometrician.

To help avoid the natural urge to maximize $\bar{R}^2$ without regard to the rest of the equation, you might find it useful to imagine the following conversation:

You: Sometimes, it seems like the best way to choose between two models is to pick the one that gives the highest $\bar{R}^2$.

Your Conscience: But that would be wrong.

You: I know that the goal of regression analysis is to obtain the best possible estimates of the true population coefficients and not to get a high $\bar{R}^2$, but my results "look better" if my fit is good.

Your Conscience: Look better to whom? It's not at all unusual to get a high $\bar{R}^2$, but then find that some of the regression coefficients have signs or magnitudes that are contrary to theoretical expectations.

You: Well, I guess I should be more concerned with the logical relevance of the explanatory variables than with the fit, huh?

Your Conscience: Right! If in this process we obtain a high $\bar{R}^2$, well and good, but if $\bar{R}^2$ is high, it doesn't mean that the model is good.

[8] For more on spurious regression and spurious correlation, see Section 12.5.

2.6 Summary

1. Ordinary Least Squares (OLS) is the most frequently used method of obtaining estimates of the regression coefficients from a set of data. OLS chooses those $\hat{\beta}$s that minimize the summed squared residuals $(\sum e_i^2)$ for a particular sample.

2. R-bar-squared $(\bar{R}^2)$ measures the percentage of the variation of Y around its mean that has been explained by a particular regression equation, adjusted for degrees of freedom. $\bar{R}^2$ increases when a variable is added to an equation only if the improvement in fit caused by the addition of the new variable more than offsets the loss of the degree of freedom that is used up in estimating the coefficient of the new variable. As a result, most researchers will automatically use $\bar{R}^2$ when evaluating the fit of their estimated regression equations.

3. Always remember that the fit of an estimated equation is only one of the measures of the overall quality of that regression. A number of other criteria, including the degree to which the estimated coefficients conform to economic theory and expectations (developed by the researcher before the data were collected) are more important than the size of $\bar{R}^2$.

EXERCISES

(The answers to the even-numbered exercises are in Appendix A.)

1. Write the meaning of each of the following terms without referring to the book (or your notes), and compare your definition with the version in the text for each:
   a. degrees of freedom (p. 53)
   b. estimate (p. 37)
   c. estimator (p. 37)
   d. multivariate regression coefficient (p. 41)
   e. Ordinary Least Squares (OLS) (p. 36)
   f. R^2 (p. 50)
   g. $\bar{R}^2$ (p. 54)
   h. total, explained, and residual sums of squares (pp. 47, 48)

2. Just as you are about to estimate a regression (due tomorrow), massive sunspots cause magnetic interference that ruins all electrically powered machines (e.g., computers). Instead of giving up and

flunking, you decide to calculate estimates from your data (on per capita income in thousands of U.S. dollars as a function of the percent of the labor force in agriculture in 10 developed countries) using methods like those used in Section 2.1 *without* a computer. Your data are:

Country	A	B	C	D	E	F	G	H	I	J
Per Capita Income	6	8	8	7	7	12	9	8	9	10
% in Agriculture	9	10	8	7	10	4	5	5	6	7

a. Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$.
b. Calculate R^2 and $\bar{R}^2$.
c. If the percent of the labor force in agriculture in another developed country was 8 percent, what level of per capita income (in thousands of U.S. dollars) would you guess that country had?

3. Consider the following two least-squares estimates of the relationship between interest rates and the federal budget deficit in the United States:

$$\text{Model A: } \hat{Y}_1 = 0.103 - 0.079X_1 \qquad R^2 = .00$$

where: Y_1 = the interest rate on Aaa corporate bonds
X_1 = the federal budget deficit as a percentage of GNP
(quarterly model: $N = 56$)

$$\text{Model T: } \hat{Y}_2 = 0.089 + 0.369X_2 + 0.887X_3 \qquad R^2 = .40$$

where: Y_2 = the interest rate on 3-month Treasury bills
X_2 = the federal budget deficit in billions of dollars
X_3 = the rate of inflation (in percent)
(quarterly model: $N = 38$)

a. What does “least-squares estimates” mean? What is being estimated? What is being squared? In what sense are the squares “least”?
b. What does it mean to have an R^2 of .00? Is it possible for an R^2 to be negative?
c. Based on economic theory, what signs would you have expected for the estimated slope coefficients of the two models?
d. Compare the two equations. Which model has estimated signs that correspond to your prior expectations? Is Model T automatically better because it has a higher R^2? If not, which model do you prefer and why?

4. Let's return to the height-weight example on page 53 and recall what happened when we added a nonsensical variable that measured the student's campus post office box number (MAIL) to the equation. The estimated equation changed from:

$$\widehat{\text{WEIGHT}} = 103.40 + 6.38\text{HEIGHT}$$

to:

$$\widehat{\text{WEIGHT}} = 102.35 + 6.36\text{HEIGHT} + 0.02\text{MAIL}$$

a. The estimated coefficient of HEIGHT changed when we added MAIL to the equation. Does that make sense? Why?

b. In theory, someone's weight has nothing to do with their campus mail box number, yet R^2 went up from .74 to .75 when MAIL was added to the equation! How is it possible that adding a nonsensical variable to an equation can increase R^2?

c. Adding the nonsensical variable to the equation decreased $\bar{R}^2$ from .73 to .72. Explain how it's possible that $\bar{R}^2$ can go down at the same time that R^2 goes up.

d. If a person's campus mail box number truly is unrelated to their weight, shouldn't the estimated coefficient of that variable equal exactly 0.00? How is it possible for a nonsensical variable to get a nonzero estimated coefficient?

5. Suppose that you have been asked to estimate a regression model to explain the number of people jogging a mile or more on the school track to help decide whether to build a second track to handle all the joggers. You collect data by living in a press box for the spring semester, and you run two possible explanatory equations:

$$\text{A: } \hat{Y} = 125.0 - 15.0X_1 - 1.0X_2 + 1.5X_3 \qquad \bar{R}^2 = .75$$
$$\text{B: } \hat{Y} = 123.0 - 14.0X_1 + 5.5X_2 - 3.7X_4 \qquad \bar{R}^2 = .73$$

where: Y = the number of joggers on a given day
X_1 = inches of rain that day
X_2 = hours of sunshine that day
X_3 = the high temperature for that day (in degrees F)
X_4 = the number of classes with term papers due the next day

a. Which of the two (admittedly hypothetical) equations do you prefer? Why?

b. How is it possible to get different estimated signs for the coefficient of the same variable using the same data?

6. What's wrong with the following kind of thinking: "I understand that R^2 is not a perfect measure of the quality of a regression equation because it always increases when a variable is added to the equation. Once we adjust for degrees of freedom by using $\bar{R}^2$, though, it seems to me that the higher the $\bar{R}^2$, the better the equation."

7. Suppose that you work in the admissions office of a college that doesn't allow prospective students to apply by using the Common Application.[9] How might you go about estimating the number of extra applications that your college would receive if it allowed the use of the Common Application? An econometric approach to this question would be to build the best possible model of the number of college applications and then to examine the estimated coefficient of a dummy variable that equaled one if the college in question allowed the use of the "common app" (and zero otherwise).

For example, if we estimate an equation using the data in Table 2.3 for high-quality coed national liberal arts colleges, we get:

$$\widehat{APPLICATION}_i = 523.3 + 2.15SIZE_i - 32.1RANK_i + 1222COMMONAPP_i \quad (2.18)$$

$$N = 49 \qquad R^2 = .724 \qquad \bar{R}^2 = .705$$

where: $APPLICATION_i$ = the number of applications received by the *i*th college in 2007

$SIZE_i$ = the total number of undergraduate students at the *i*th college in 2006

$RANK_i$ = the *U.S. News*[10] rank of the *i*th college (1 = best) in 2006

$COMMONAPP_i$ = a dummy variable equal to 1 if the *i*th college allowed the use of the Common Application in 2007 and 0 otherwise.

[9] The Common Application is a computerized application form that allows high school students to apply to a number of different colleges and universities using the same basic data. For more information, go to www.commonap.org.

[10] U.S. News and World Report Staff, *U.S. News Ultimate College Guide*. Naperville, Illinois: Sourcebooks, Inc., 2006–2008.

Table 2.3 Data for the College Application Example

COLLEGE	APPLICATION	COMMONAPP	RANK	SIZE
Amherst College	6680	1	2	1648
Bard College	4980	1	36	1641
Bates College	4434	1	23	1744
Bowdoin College	5961	1	7	1726
Bucknell University	8934	1	29	3529
Carleton College	4840	1	6	1966
Centre College	2159	1	44	1144
Claremont McKenna College	4140	1	12	1152
Colby College	4679	1	20	1865
Colgate University	8759	1	16	2754
College of the Holy Cross	7066	1	32	2790
Colorado College	4826	1	26	1939
Connecticut College	4742	1	39	1802
Davidson College	3992	1	10	1667
Denison University	5196	1	48	2234
DePauw University	3624	1	48	2294
Dickinson College	5844	1	41	2372
Franklin and Marshall College	5018	1	41	1984
Furman University	3879	1	41	2648
Gettysburg College	6126	1	45	2511
Grinnell College	3077	1	14	1556
Hamilton College	4962	1	17	1802
Harvey Mudd College	2493	1	14	729
Haverford College	3492	1	9	1168
Kenyon College	4626	1	32	1630
Lafayette College	6364	1	30	2322
Lawrence University	2599	1	53	1409
Macalester College	4967	1	24	1884
Middlebury College	7180	1	5	2363
Oberlin College	7014	1	22	2744
Occidental College	5275	1	36	1783
Pitzer College	3748	1	51	918
Pomona College	5907	1	7	1545
Reed College	3365	1	53	1365
Rhodes College	3709	1	45	1662
Sewanee-University of the South	2424	0	34	1498
Skidmore College	6768	1	48	2537
St. Lawrence University	4645	0	57	2148

(continued)

Table 2.3 (*continued*)

COLLEGE	APPLICATION	COMMONAPP	RANK	SIZE
St. Olaf College	4058	0	55	2984
Swarthmore College	5242	1	3	1477
Trinity College	5950	1	30	2183
Union College	4837	1	39	2178
University of Richmond	6649	1	34	2804
Vassar College	6393	1	12	2382
Washington and Lee University	3719	1	17	1749
Wesleyan University	7750	1	10	2798
Wheaton College	2160	1	55	1548
Whitman College	2892	1	36	1406
Williams College	6478	1	1	2820

Sources: U.S. News & World Report Staff, *U.S. News Ultimate College Guide*, Naperville, IL: Sourcebooks, Inc. 2006–2008.

Datafile = COLLEGE2 (Note that some colleges tied for the same rank.)

a. Take a look at the signs of each of the three estimated regression coefficients. Are they what you would have expected? Explain.

b. Carefully state the real-world meaning of the coefficients of SIZE and RANK. Does the fact that the coefficient of RANK is 15 times bigger (in absolute value) than the coefficient of SIZE mean that the ranking of a college is 15 times more important than the size of that college in terms of explaining the number of applications to that college? Why or why not?

c. Now carefully state the real-world meaning of the coefficient of COMMONAPP. Does this prove that 1,222 more students would apply if your college decided to allow the Common Application? Explain. (*Hint:* There are at least two good answers to this question. Can you get them both?)

d. To get some experience with your computer's regression software, use the data in Table 2.3 to estimate Equation 2.18. Do you get the same results?

e. Now use the same data and estimate Equation 2.18 again without the COMMONAPP variable. What is the new $\bar{R}^2$? Does $\bar{R}^2$ go up or down when you drop the variable? What, if anything, does this change tell you about whether COMMONAPP belongs in the equation?

2.7 Appendix: Econometric Lab #1

Throughout the text, you'll encounter "econometric labs" aimed at helping you get experience with the chapter topics.[11] We urge you to complete these labs, even if they aren't assigned as homework by your professor. Working through the labs on your own will make future econometric work much easier to understand and complete.

Our first lab is an exercise in simple regression analysis. You will collect a data set and then calculate regression coefficients and goodness of fit measures on your own (using Stata or another econometric software package to run the regression). Your goal is to estimate the relationship between the dependent variable WEIGHT and the independent variable HEIGHT, using your own sample, not the sample from the book.

Step 1: Gather the Data

Ask five students of your gender how tall they are, in inches, and how much they weigh, in pounds. Also report your own height and weight. Do not include names in the data. Record the variable HEIGHT in inches above five feet and the variable WEIGHT in pounds. Enter the data in an Excel spreadsheet (or directly into Stata) according to the following instructions. In the first row, type the first person's height in inches above five feet, the first person's weight in pounds, and the first person's gender—1 if male, 0 if female.

For the data, be sure to enter only numbers, not words such as pounds, inches, or male. Otherwise, Stata will get irritated and refuse to run the regression! In the height column, enter the height in inches above five feet. Enter the weight in pounds. If the observation is male, enter 1 in the MALE column. If female, enter 0.

After you enter all the data, your file might look like this:

HEIGHT	WEIGHT	MALE
4	127	0
9	152	0
6	130	0
2	130	0
6	112	0
3	119	0

[11] These labs are simplified versions of labs designed by Bruce Johnson for use at Centre College. Instructors should consult the instructor's manual at http://www.pearson.com/studenmund for the answers to the labs and for suggestions on how best to use these labs in a classroom setting.

Now carry out the following tasks and answer the questions in order, using the data you collected.

Step 2: Calculate Summary Statistics

Use Stata to compute the summary statistics. What is the mean? Also look at the minimum, the maximum, and the sample size.

Step 3: Run the Regression

Run the regression in Stata. Can you find the estimated βs? Print the regression results if you are doing this lab as a class assignment.

Step 4: Interpret the Estimated Coefficients

State the precise meaning of the slope coefficient. Now compare your estimated coefficients with those in Equation 2.6. Are your results the same? If not, do you understand why they're different?

Step 5: Interpret $\bar{R}^2$

Can you find $\bar{R}^2$ in your results? What is it? State precisely the meaning of the $\bar{R}^2$ statistic.

Step 6: Estimate a Second Equation

Now add MALE to your equation as a second independent variable and estimate it again. Do you see a problem? Explain. Does this imply that gender has no relationship to weight?